NAND 的替代技术——3D 闪存

作者：Andrew J. Walker
Schiltron Corporation

传统的 NAND Flash 根本不能被视为一种“通用”存储器。的确，其特殊的线性串列 MOSFET 存储器结构限制了它在那些相对较慢存储时间就已经足够好的领域的应用。东芝公司(Toshiba)于 1988 年在国际电子装置会议(International Electron Device Meeting)上发表了有关 NAND Flash 的首篇学术论文，那时的 NAND Flash 是由在两个使用 1um 设计原理的存取设备之间排成一列的 8 个浮栅 MOSFET 组成。從那时以后我们的技术一直不断进步。写这篇文章时，最先进的产品使用的是半节距约为 30nm 的装置。自线性缩放出现 20 年来，许多创新一直推动着这种技术的快速发展。

NAND Flash如此流行的原因包括较低的每位成本和个人存储需求的显著上升。NAND的单元尺寸接近 $4F^2$（F是该方法的最小形体尺寸）的理论极限。线性串列结构使得节距可以根据朝着一个方向的无接触栅间隔与朝着另一方向的无接触场氧化间隔确定。随后两个串列间可以共享位线接触，而且多个串列可以共享源极。

在我写这篇文章时，闪存記憶体高峰会(Flash Memory Summit)刚刚在圣塔克拉拉(Santa Clara)闭幕。SanDisk 公司创始人兼首席执行官 Eli Harari 在会上发表了一些有趣的评论。除了承认该行业将见证“NAND 可微缩性发展速度的放缓”之外，他还指出 3-D Flash 是 NAND 潜在的可行替代技术。具体来说，“如果材料方面取得突破”，SanDisk 和东芝公司共同研究的 3-D 技术将能够取代 NAND。此外，他还表示“向 3-D 技术过渡还需要数年时间”。

因此有何原因 3-DFlash 被认为是 NAND 潜在的替代技术呢？又有哪些技术可以取代 NAND Flash 的王者地位呢？本文中的第二部分讨论了 NAND Flash 可微缩性发展速度将会放缓的原因，并对 NAND Flash 的各种潜在替代技术进行了分析。第三部分对 3-D Flash 和 2-D Flash 的成本进行了比较。第四部分对 3-D 产品的发展以及标准 2-D NAND Flash 优势的逐渐消失进行了一些预测。

NAND 可微缩性和 3-D 技术

Kirk Prall 在微米(Micron)上的著作是我所知道的探讨有关 NAND 微缩范围在 30nm 以内的问题的最佳文章。在这篇文章中，他指出浮栅结构是这些问题的主要原因。尤其是，浮栅周围的电容耦合干扰可为任一特定的浮栅带来特定模式电压，这使得在每个单元存储多位信息的能力有所降低。

随着可微缩性发展速度的持续下降，或许需要利用极端远紫外光刻以及它涉及的所有生产架构变化，因此经济因素也是导致 NAND 可微缩性发展速度放缓的原因。

以下是已经提出的向 3-D Flash 技术过渡的多种方法。

A. 电阻变化法

这里可分为几种方法。对于 3-D 结构中的最小单元而言，大多数（如果不是全部）可以遵循从这个参考书目第 28 页复制而来的图 1 中所示的方法。请注意导引元件（最可能是一个二极管），因为这是一个二端装置，并且可以在很小的区域中制造。

电阻变化法的第一个例子是相变存储器(Phase Change Memory，PCM)，该存储器涉及对常以 Ge2-Sb2-Te5(GST)形式存在的硫属化物原料的使用。可程序化的有阻力状态取决于非晶相和（聚乙烯）晶相之间的可逆变化。将低电阻改为高电阻状态的复位电流通常约为几百 μA。

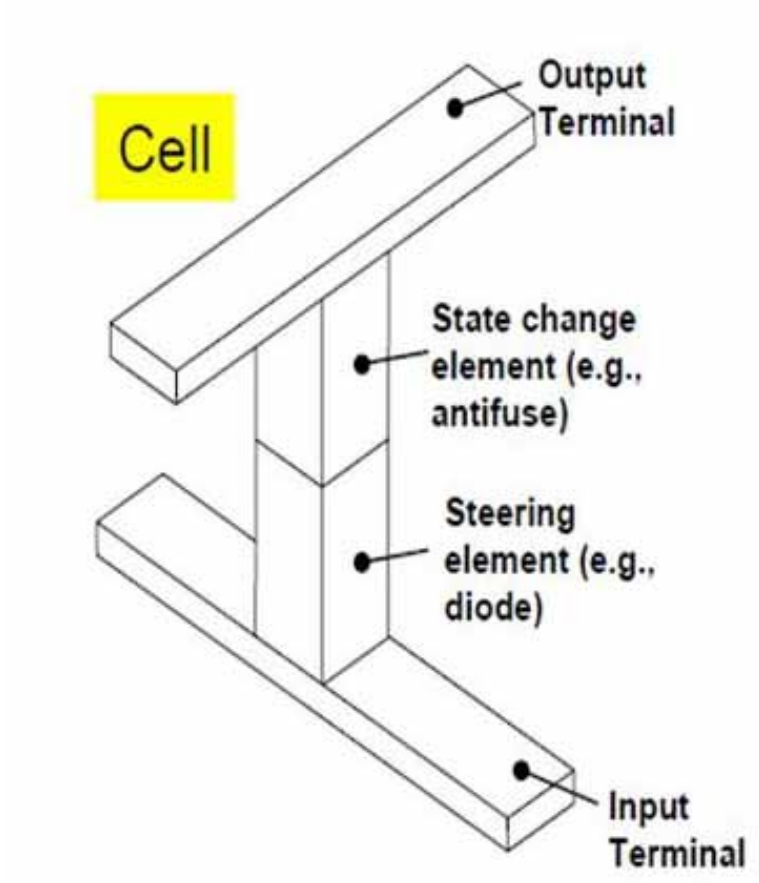

图 1

这个类别中的另一种方法涉及对一些钙钛矿原料的使用。对于 PCM 方法而言，将状态从低电阻改为高电阻似乎至少需要 100 μA。

第三种方法是将纯金属氧化物用作切换材料。正如上文提到的，复位电流至少需要 100 μA，尽管最近出现了复位电流可低于 100 μA 的利好数据。

还可以通过其它有趣的方法实现电阻切换，包括报导中所使用的编程电流低至 1 μA 的固态电解质。

有关切换材料的文献指出，在实现高密度 3-D 技术之前依然面临一些挑战，也就是说需要在足够低的工艺热预算中整合高电流驱动导引装置，并需要将复位电流降至能够出现 NAND Flash 中已经存在的大量程序并行的水平。另一项考虑或许就是需要将复位电流降至能够将精选晶体管整合进 3-D 存储器的水平。目前基于多晶硅或纳米硅的薄膜晶体管技术不能满足上述切换材料的现有需求。

B. 浮栅和采用水平存储技术的 NAND 电荷捕获闪存存储器

实现 3-D Flash 技术的第二大类方法涉及对一些以串接方式组成的晶体管的使用。最明显的是对现有的 NAND 浮栅结构进行堆叠。之后它便拥有与普通 NAND 浮栅相同的横向可微缩性限制，并预计将会在约 30nm 的半节距处遇到相同的困难。迄今为止公布的所有其它 3-D 串接技术都涉及对利用氮化硅来代替浮栅的电荷捕获闪存(CTF)方法的使用。

欲了解 CTF NAND 方法所面临的挑战，请参考图 2。

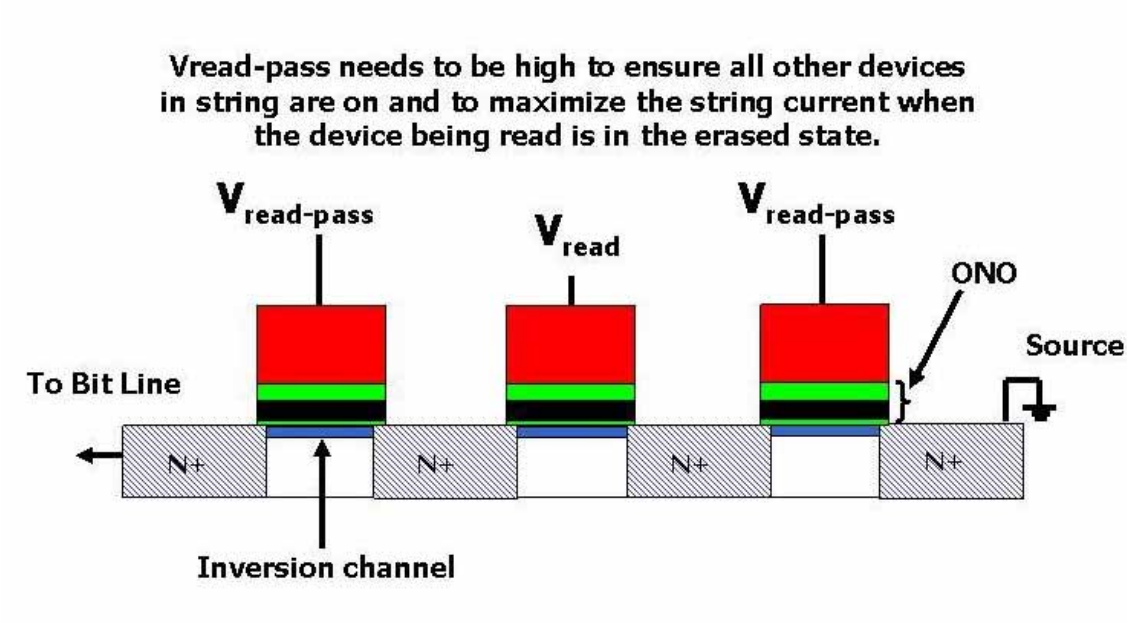

图 2

第一项挑战是确保在一个单元需要被读取时，所有传输单元都在运作。因此，读取传输电压需要比有可能最高的编程阈值电压和电压容限之和还要高。如果电压容限小，那么所有编程传输单元都具有高阻抗性，能够在正在被读取的单元被擦除（低阈值电压）时实现较低的串电流。第二项挑战是在所有传输单元均处于最高编程状态以及被读取的单元被擦除时确保“最糟情况”下有足够大的串电流。这也需要很高的读取传输电压。第三项挑战便是确保在操作时决不会出现阈值电压高于读取传输电压的情况。否则，传输单元将在连接正被读取的单元或对阵列末端进行编程时会断开。同样，如果读取传输电压保持不变，那么耐久性周期中阈值电压的升高会导致串电流的大量减少，或导致读取传输电压上升以及随之按指数增加的读取传输干扰。NAND Flash 通过采用增加一个阵列中单元的数量这种常用方法来提高面

积效率。这样，位线以上的区域可以相互接触，源极也可以分摊到更多单元。如果读取传输电压不能大幅提高，那么这种方法将降低最糟情况下的串电流。

在这种传输电压下，薄隧道氧化物硅氧化氮氧化硅（SONOS）设备可实现软件编程，该解决方案已将隧道氧化物层从最初的 2.5nm 左右增厚到约 5nm。然而，这却产生了很高的编程和擦除电压、来自传输电压的不可避免的指数阈值电压效应以及随之产生的 TANOS 结构有限的耐久性。由于薄膜晶体管很低的载流子迁移率，薄膜晶体管串接的情况尤为棘手。

三星是以 TANOS 形式存在的 3-D 版 CTF NAND 的主要倡导者。然而，即使凭借相对较厚的 TANOS 栅极介质堆栈，通过以更高的编程和擦除电压为代价来完全消除传输干扰也是无法实现的。

Schiltron 公司率先开发出了一种可以完全消除传输干扰并依然采用薄的 ONO 栅极介质堆栈的方法，详见图 3。

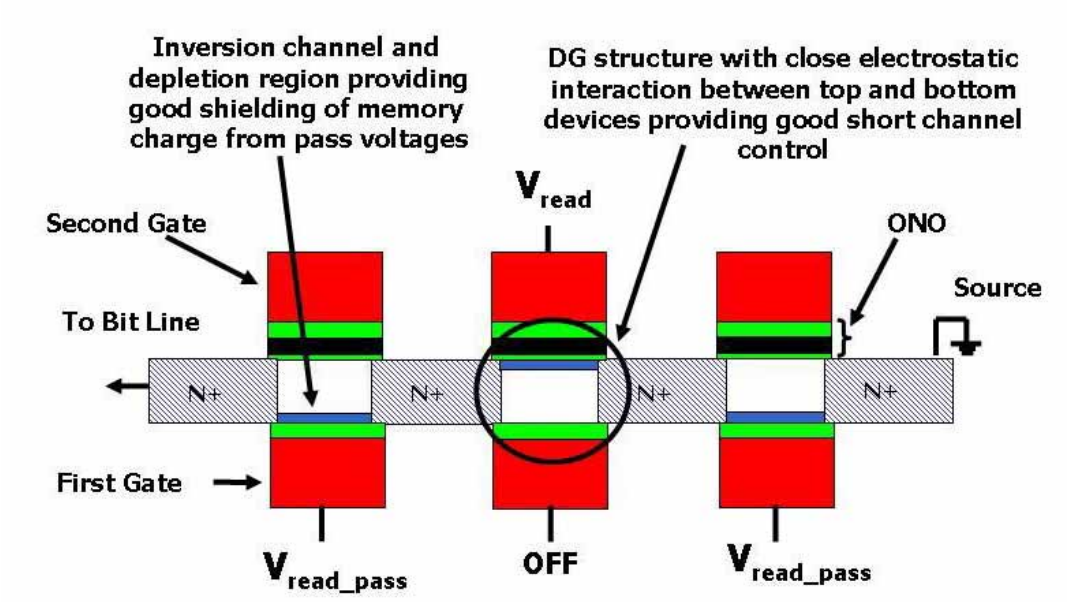

图 3

在此，串联存储器的存取可通过在低级非存储器中形成一个反型沟道来实现。这种反型沟道及其相关耗尽区为存储介质中所捕获的电荷提供了高水平电荷保护，免受应用于这些底部存取器的传输电压的干扰。此外，这种双栅结构是公认的良好横向可微缩性方法，它通过使顶部和底部设备之间实现密切的静电相互作用来消除短通道效应。去年国际电子装置会议的一篇论文[1]描述了这种基于硅的最小尺寸 TFT（长和宽均低于 50nm），其阵列由多达 64 个单元

组成，在最糟的情况下串电流也能达到上百的 nA。其它优势还包括采用了标准材料，具有标准程序和擦除功能，以及足以在彼此顶部实现多层堆叠的低工艺温度预算。将这一结果与其它需要突破性材料的方法进行对比，你会发现十分有趣。

C．采用垂直存储技术的 NAND 电荷捕获闪存存储器

串联 CTF 设备的一项重要技术就是垂直 NAND。与横向 3-D NAND 技术相比，该技术旨在使每位成本降到更低。东芝的 BiCS（三维层叠技术）技术由介质隔离导电板上的蚀刻孔构成。这些蚀刻孔之后将由存储介质堆栈和硅通道填充，最终成为一种垂直 CTF NAND 结构，其中每个串联装置在该通道周围都有一个栅极。然而，一些重要挑战依然存在。例如，以上所述的同样的传输干扰分析也适用。此外，横向可微缩性与需要同时采用 CTF 介质堆栈和硅通道来填充该蚀刻孔有关，将横向半节距大概限制在 55nm 以上。最后，通过延长每个垂直 NAND 阵列来增加存储密度的方法，不仅增加了传输干扰，还降低了最糟情况下的串电流。的确，密度每增加一倍，最糟情况下的串电流就会减半。由于这些装置的通道是多晶硅，因此，随着密度的增加，最糟情况下的串电流可能会迅速降到不可读取的低值。此外，事实是利用超薄 CTF 介质堆栈来构建一个小单元不仅会对传输干扰产生重大影响，并且还会影响保留率。尽管如此，BiCS 技术依然是降低成本的一种积极尝试。

3-D Flash 的成本优势

首先解释 3-DFlash。这是从基片逐个堆叠起来的闪存单元，基片可能含有驱动电路。堆叠是通过材料沉积与沉积材料中构建的闪存单元来实现。这种方法被称为单片集成(monolithic)，与芯片堆叠等相比是制造 3-D 闪存芯片最具成本优势的方法。该方法最近获得了极高的评价。

堆叠闪存单元的成本优势并不是特别明显，这是因为为了获得一种具备一定存储能力且总体尺寸较小的芯片，该工艺的成本正逐渐增加。工艺复杂性的加深也可能影响产量，也会降低成本优势。我们不能简单地说选择芯片尺寸较小的 3-D 技术就一定具有优势，真正找到成本优势解决方案才是关键。

正是由于这个原因，我开发了一种综合型单片 3-D 成本模式，并且已于近期推出[2]。我将在此对这项方法和结论加以总结。

在单片 3-D 中，有一个因素降低了成本，即在固定存储容量的条件下，每晶圆可提供更多芯片。很多支持 3-D 的人士十分注重这一点。然而，提高工艺的复杂性（需要额外的工艺步骤）和产量这两个因素增加了成本。而这些因素会有效的 3-D 闪存芯片的总成本产生怎样的影响呢？

我的成本模型论文将所有这些要素结合成一个成本方程式，通过与容量相同的 2-D 闪存芯片的成本对比，来分析 3-D 闪存芯片的成本。图 4 显示该篇论文中的一个例子，2-D 芯片采用了半节距为 25nm 的 NAND，而 3-D 芯片采用的是 32nm 半节距的 NAND。参数 Z 代表的是从一个节点到下一个更先进节点的设备成本的倍数。在此，它有所不同，可作为一个显示不断上涨的 2-D NAND 制作成本的数字。

通过这一分析可以发现以下几个有趣的特点。

首先，3-D 存储单元层的数量达到最佳值可使相关成本达到最低。请看如下说明。设想在一个 2-D NAND 闪存芯片里，即使存储单元被置于 3-D 堆栈中，控制电路仍将保留在基片中。被存储单元占用的 2-D 芯片的面积通常被成为阵列效率。剩余的空间通常占 2-D 芯片总面积的四分之一左右。因此，如果这个电路被置于存储堆栈之下，3-D 芯片面积就达到最小值。为了实现这一最小面积，将需要约 4 个存储层，因为层数每增加一倍，阵列面积就将减少一半。存储层数达到 4 将使该面积几乎与控制电路留下的面积相等。

第二，3-D 堆叠的成本并未像有些人所说的降低了十倍。相反，一旦难以做到横向可微缩性，自然可以用 3-D 堆叠来替代。于是，通过 3-D 堆叠实现成本的进一步下降，与横向可微缩性的效果大致相同。

第三，将所有存储单元从基片移至 3-D 堆栈中是更好的选择。实际上，通过增加层数是无法实现最低成本的，而这样所耗成本始终比将所有存储单元置于基片之外的成本更高。

第四，控制电路融入 3-D 堆栈技术将更有利于降低成本，甚至更小的芯片尺寸可以实现。

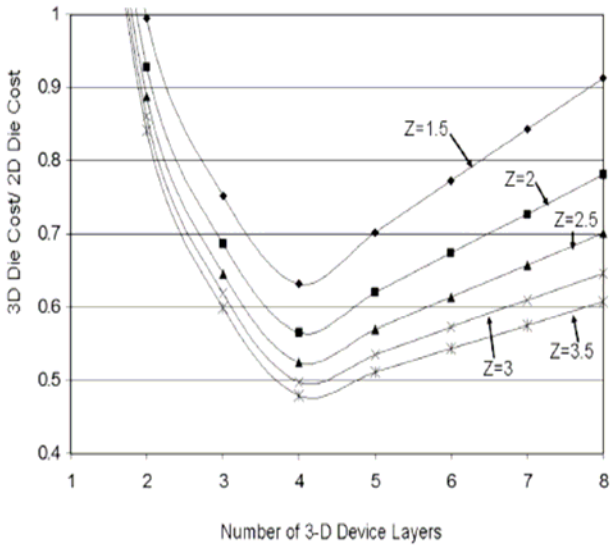

Fig. 5. Ratio of 3-D to 2-D die costs as a function of the number of 3-D device layers added for a fixed capacity memory (64 Gbit in this case) with Z as a parameter. 2-D uses advanced 25-nm half-pitch while 3-D uses more mature 32-nm half-pitch. $D_0 = 0.1$ cm^2, $MLC = 2$ and $AE_0 = 75\%$. The same base wafer and adder costs were used as in Fig. 4.

图 4（取自成本模型参考）

预测

曾经有人说，一个人千万不要预测，尤其是对未来的预测。不过，在这里，我将做出预测（当然，我无法保证未来）。

首先，我预测，2-D NAND 将战略性地撤出某些细分市场并无法再满足这些市场的要求。原因是，现在正在执行一些功能上的折衷方案并将使每位成本继续降低。在这里，我们已经能够看出可靠性上的降低（耐久性和循环后的保留率），为了争取先进节点上每单元的多个比特数。显然，许多市场将无法避免这一情况，并将在至少一段时间内继续采用 2-D NAND。

第二，我预测，3-D 闪存存储器将逐步占领一些高密度闪存细分市场。进展可能相对较慢，但 3-D 闪存存储器将必须在市场上获得认可。我想起来另外一个例子，“ 日本的摩托车方法 ”。我很清楚地记得，日本制造商生产的 50cc 小型摩托车在英国上市，而当时英国正是摩托车行业的统治者。专家表示，这些小型摩托车是不可能取代英国在这个领域的最高地位。然而，我们看到一切都发生了。小处入手，逐步占领市场。

第三，首款包含 3-D Flash 的产品距今还有三年的时间上市。问题是：哪种 3-D 技术以及哪些细分市场？在这里，我或许可以不谈客观事实，说出自己的看法。将占据领导地位的 3-D 闪存技术将仍然处于半导体行业规则的范围之内，即 “ 尽可能少地改变 ”。在这方面，3-D 技术利用现有的材料和工具，尽可能地接近于现有的编程和擦除功能，从而将最终取得主导地位。